Early Identification of Infant Brain Abnormalities via Efficient 3-D CNN Screening of MRI Scans

Janam Chahal
The Harker School
San Jose, California
e-mail: 26janamc@students.harker.org

Kamalashree S
AIClub
Mountain View, California
e-mail: kamalashree.s@aiclub.ai

Sindhu Ghanta
AIClub
Mountain View, California
e-mail: sindhu.ghanta@aiclub.ai

***Abstract*—Early identification of brain abnormalities in infancy is critical for optimizing neurodevelopmental outcomes. However, pediatric Magnetic Resonance Imaging (MRI) interpretation remains time-intensive and resource-constrained, hindering timely medical response. Unlike prior 2-D deep-learning approaches targeting narrow tasks, in this study, we investigate whether 3-D convolutional neural networks can assist pediatric care by rapidly screening T2-weighted MRIs for structural abnormalities. We curated a cohort of 833 MRI scans from patients aged 0–36 months, standardized to a $96 \times 96 \times 96$ grid. We trained and evaluated two architectures, DenseNet-121 and DenseNet-264, adapted for 3-D volumetric input. On a strictly held-out test set, DenseNet-121 achieved an accuracy of 75.90%, outperforming the deeper DenseNet-264 (74.70%). While DenseNet-121 demonstrated higher sensitivity, which is crucial for screening tasks, DenseNet-264 reduced false positives at the cost of sensitivity. Our findings demonstrate the feasibility of end-to-end 3-D classification for early pediatric MRI, establishing a baseline for future work and eventually help accelerate pediatric MRI triage and enable earlier clinical intervention. The results highlight that increased model depth does not intrinsically yield better generalization on modest pediatric datasets. Future clinical integration requires balancing of the architectural capacity and specificity.**

Keywords— ***Pediatric MRI***; ***3-D Convolutional Neural Networks***; ***Infant Brain Abnormalities***; ***Medical Image Classification***;

I. Introduction

In pediatric care, early recognition of structural brain abnormalities is critical because neurodevelopment in the first years of life is rapid and highly plastic. Timely identification enables earlier intervention and better long-term outcomes. Consistent with this, professional guidance emphasizes developmental surveillance at every well-child visit and formal screening in the first three years of life [1]. MRI is often the preferred modality when imaging is indicated, owing to its sensitivity for malformations, white-matter/myelination disorders, and migrational anomalies relevant to developmental delay [2][3]. Moreover, early MRI findings can predict later neurodevelopmental performance in high-risk infants, underscoring the value of prompt and reliable image interpretation [4]. Yet, pediatric MRI remains resource-intensive. Young children frequently require deep sedation or anesthesia to minimize motion, which adds time, cost, and safety considerations to already busy services [5][6]. At the health-system level, rising imaging volumes and a persistent radiology workforce shortfall have stretched reporting capacity, contributing to delays in delivering actionable results—especially for complex modalities like MRI [7].

Manual interpretation of pediatric brain MRI is painstaking and time-intensive. A single study comprises hundreds of slices that must be reviewed systematically for age-appropriate myelination patterns, malformations of cortical development, white-matter signal changes, and incidental findings. Turnaround times can lengthen when studies require subspecialty input, when motion artifacts necessitate repeat sequences, or when service volumes spike. Since interpretation is a task for trained experts in the field, AI systems are well positioned as assistive tools that triage, prioritize, and standardize measurements. These systems are never used as replacements for clinical judgment [8]. In this context, a fast and reliable screening model that flags potentially abnormal studies could help reduce time-to-report for the most urgent cases while maintaining human oversight [9].

These pressures have accelerated interest in Artificial Intelligence (AI) as an assistive tool in imaging. Meta-analyses indicate that deep learning systems can achieve diagnostic performance comparable to clinicians under study conditions, while also highlighting the need for rigorous external validation and careful study design [10]. Despite growing interest in AI for pediatric neuroimaging, key gaps remain. Most prior studies focus on narrow clinical tasks and rely on 2-D slice-based models that inadequately capture 3-D anatomical context, particularly important in rapidly developing infant brains. Very few works evaluate volumetric deep learning approaches in the 0–36-month age range, where rapid maturation and heterogeneous acquisition protocols pose unique challenges. Dataset sizes are often small, performance is rarely assessed on rigorously held-out cohorts, and the impact of model depth under limited pediatric data is largely unexplored. For volumetric data such as brain MRI, 3-D convolutional networks explicitly model inter-slice anatomical context and can outperform 2-D, slice-based approaches across several applications [11].

In this work, we formulate early detection of pediatric brain abnormalities as a supervised 3-D volumetric classification problem on T2-weighted MRI. We assemble 833 scans from patients aged 0–36 months acquired on a single scanner (under varying protocols), binarize labels into Normal vs. Not healthy from radiology reports, and standardize each volume to a 96×96×96 grid for efficient batching. We then compare two densely connected architectures, (a) DenseNet-121 and (b) DenseNet-264, adapted to 3-D convolutions. Our study is designed to (i) quantify the feasibility and headroom of

end-to-end 3-D classification in this age range, (ii) assess depth/capacity trade-offs under limited pediatric data, and (iii) report clinically meaningful metrics such as accuracy and confusion matrix on a strictly held-out test set. The broader goal is to evaluate whether such a model can pragmatically support pediatric neuroradiology workflows by accelerating case prioritization while preserving expert oversight [12].

The remainder of this paper is organized as follows. Section II reviews prior work in medical image analysis and pediatric neuroimaging. Section III details the dataset, preprocessing steps, model architectures, and training procedures, while Section IV presents the experimental results. Section V provides a critical evaluation of model performance and clinical relevance, and Section VI concludes with future directions for improving pediatric MRI classification.

II. Related work

Deep learning has reshaped medical image analysis across modalities and tasks such as classification, detection, segmentation, registration, and even reconstruction. This is accomplished by replacing hand-engineered features with end-to-end representation learning documented in comprehensive surveys of the field [13]. Meta-analytic evidence further suggests that, under study conditions, diagnostic performance of deep models can match that of health-care professionals, while also underscoring persistent gaps around external validation and study quality [10].

Within neuroimaging, volumetric MRI has catalyzed 3-D convolutional approaches that explicitly model inter-slice context; seminal architectures such as 3D U-Net and multiscale 3-D Convolutional Neural Networks (CNN) (with CRF post-processing) established strong baselines for brain lesion segmentation [14]. Subsequent reviews report robust performance of 3-D CNNs across segmentation and classification problems in MRI and other modalities, including neurodegenerative disease classification [15]. Community benchmarks like the Brain Tumor Segmentation (BraTS) challenge have further accelerated progress by standardizing datasets and evaluation protocols and by tracking gains across successive competition years [16].

Beyond general surveys of medical imaging AI, pediatric neuroimaging presents distinct challenges. There are rapid, non-linear changes in tissue contrast due to myelination, smaller brain size, and a higher prevalence of congenital or developmental anomalies relative to adult cohorts. Much of the mature literature and public benchmarks, such as BraTS centers on adult pathologies and segmentation tasks, leaving pediatric classification comparatively underexplored. Prior pediatric efforts have often targeted specific conditions (e.g., age estimation/myelination staging, neonatal white-matter injury, posterior fossa tumors) are relied on 2-D slice-based models with ensembling across slices, which can dilute cross-slice anatomical context [17]–[19]. Although effective in narrow settings, these approaches cannot fully capture the subtle, spatially distributed patterns present in early-life T2 MRI and are further limited by the rapid anatomical changes and small brain structures characteristic of the 0–36 month age range.

Methodologically, 3-D CNNs capture volumetric dependencies that are critical in pediatric brains where subtle, spatially distributed patterns carry diagnostic signal. Dense connectivity (as in DenseNet) promotes feature reuse and stabilizes optimization for deeper networks—properties desirable when labeled pediatric datasets are modest in size [20]. While 3-D DenseNet variants have shown promise in adult neuroimaging and other modalities, there remain relatively few end-to-end studies on full-volume 3-D classification of pediatric T2 MRI in the first three years of life. Our work contributes to this gap by training and evaluating 3-D DenseNet-121 and DenseNet-264 on a curated 0–36-month cohort with standardized preprocessing and a class-stratified held-out test set, providing a reproducible baseline for future pediatric studies.

Unlike prior pediatric studies that rely primarily on 2-D slice-based aggregation models, this work implements full-volume 3-D DenseNet architectures for binary abnormality screening in infants aged 0–36 months. To our knowledge, few studies have systematically evaluated depth-related generalization trade-offs in 3-D CNNs under limited pediatric data conditions using a strictly held-out cohort.

III. Methodology

This section describes the dataset, preprocessing pipeline, model architectures, and evaluation protocol used to assess three-dimensional (3-D) DenseNet classifiers for early pediatric MRI abnormality screening.

A. Dataset and Data Source

We curated a cohort of 833 pediatric T2-weighted brain Magnetic Resonance Imaging (MRI) volumes from patients aged 0–36 months. All scans were acquired at a single institution using the same scanner, though imaging protocols varied across examinations. Of the 833 cases, 565 were labeled as Normal and 268 as Not healthy based on radiologist reports.

Each MRI volume was paired with structured metadata (`meta.csv`) containing an `image_id` and corresponding radiologist-derived diagnosis. Records lacking a corresponding T2-weighted volume were excluded following a file-system audit keyed by `image_id`. For supervised learning, diagnoses were binarized as *Normal* when no pathology was reported and *Not healthy* otherwise.

All labels were derived from clinical reports rather than manual re-annotation. Demographic variables such as gender distribution were not available in the metadata and therefore were not incorporated into modeling.

B. Data Pre-processing

All MRIs were originally stored in NIfTI (`.nii.gz`) format. For compatibility with the training framework, volumes were converted to PyTorch tensor (`.pt`) format.

Preprocessing consisted of the following standardized steps applied uniformly across all scans:

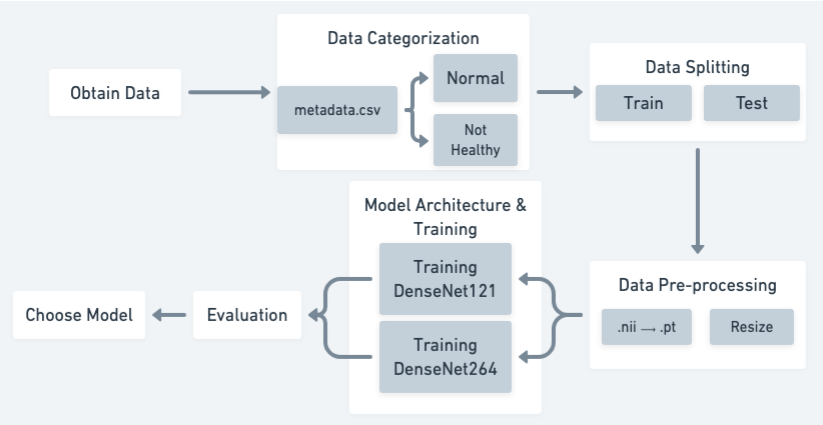


Figure 1. Flowchart of Data Processing and Machine Learning Stages

- Canonical reorientation to ensure consistent anatomical alignment,
- Intensity clipping to robust percentile ranges to reduce the influence of outliers,
- Per-volume z-score normalization,
- Uniform resampling to a fixed grid of $96 \times 96 \times 96$ voxels.

Resampling ensured consistent input dimensions across subjects and enabled efficient batched training of 3-D convolutional models.

The overall data preparation and modeling workflow is summarized in Figure 1. Notably, preprocessing was performed prior to dataset splitting to ensure consistent transformations across all samples.

C. Model Architectures

Two three-dimensional convolutional neural network (CNN) classifiers were implemented based on DenseNet-121 and DenseNet-264 architectures. All two-dimensional convolutional, pooling, and normalization operations were replaced with their three-dimensional counterparts. The initial convolution layer was adapted to accept single-channel T2-weighted input volumes.

Dense connectivity promotes feature reuse and improves gradient flow across depth, while transition layers with compression constrain parameter growth. These properties are advantageous when training deep networks on modest-sized pediatric datasets characterized by anatomical heterogeneity.

D. Model Evaluation

To obtain an unbiased estimate of generalization, we employed a stratified hold-out validation strategy. Twenty percent of cases from each class (Normal and Not healthy) were reserved as a strictly held-out test set. This cohort remained completely unseen during model training and hyperparameter tuning.

The remaining 80% of the data were used for model development, with a validation subset carved from the training portion for hyperparameter selection.

Models produced a single sigmoid output and were trained using weighted binary cross-entropy loss to mitigate class imbalance. Training employed mini-batches with on-the-fly three-dimensional data augmentation.

A coarse hyperparameter sweep explored learning rates in the range $[10^{-6}, 5 \times 10^{-2}]$ and training durations between 10 and 50 epochs. Three-dimensional DenseNet variants are known to be sensitive to optimizer step size; excessively large learning rates can cause unstable optimization, whereas excessively small values may hinder convergence. The selected range therefore spans conservative to moderately aggressive update regimes.

The upper bound of 50 training epochs reflects the tendency of high-capacity 3-D models to overfit under limited data conditions. Early stopping was guided by validation performance. The final configuration for each architecture was selected based on validation accuracy, and all performance metrics reported in Section IV were computed exclusively on the untouched test set.

Evaluation metrics included accuracy, precision, recall (sensitivity), F1-score, and confusion matrix counts.

IV. Results

A. Hyperparameter Tuning

A coarse grid search was performed over learning rates $\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}\}$ and training durations of 10, 20, 30, 40, and 50 epochs for both architectures. For DenseNet-121, validation accuracy ranged roughly from 60% to 80% across the grid. Performance improved consistently as training progressed from 10 to 50 epochs at moderate learning rates, with the best validation accuracy of 80.60% obtained at 50 epochs and a learning rate of 10^{-4}, as shown in Table I.

For the deeper 3-D DenseNet-264, validation accuracy were slightly lower on average, and there was a strong dependence on conservative learning rates. Table II displays that the best validation accuracy obtained was 78.11% at 40 epochs and a learning rate of 10^{-5}. Similar to DenseNet-121, intermediate learning rates and moderate training durations (20–40 epochs)

TABLE I. VALIDATION ACCURACIES OBTAINED DURING HYPERPARAMETER TUNING FOR DENSENET121

Epochs / Learning Rate	0.000001	0.00001	**0.0001**	0.001	0.005	0.01	0.05
10	64.68%	74.13%	71.64%	76.12%	78.11%	66.67%	71.64%
20	71.64%	75.12%	67.16%	74.63%	76.12%	71.64%	68.66%
30	73.13%	72.14%	70.65%	77.61%	74.13%	75.12%	68.66%
40	75.12%	75.62%	71.64%	74.13%	77.11%	60.70%	66.67%
50	71.64%	80.10%	**80.60%**	73.13%	74.63%	66.17%	76.12%

tended to perform well, while large learning rates (10^{-2}, 5×10^{-2}) led to earlier saturation or degradation, showing that the higher capacity model is more susceptible to overfitting and unstable updates when learning rates are too high.

Overall, the grid search indicates that (i) both architectures benefit from training beyond 20 epochs, (ii) DenseNet-121 is slightly more forgiving to a range of learning rates, and (iii) DenseNet-264 yields its best performance under more conservative optimization settings.

TABLE II. VALIDATION ACCURACIES OBTAINED DURING HYPERPARAMETER TUNING FOR DENSENET264

Epochs / Learning Rate	0.000001	**0.00001**	0.0001	0.001	0.005	0.01	0.05
10	71.14%	73.13%	73.63%	71.14%	71.64%	56.72%	65.67%
20	73.63%	76.62%	73.63%	75.12%	74.13%	69.15%	63.68%
30	72.64%	66.67%	69.65%	72.64%	76.12%	73.63%	72.64%
40	70.15%	**78.11%**	77.11%	72.64%	65.17%	67.16%	68.66%
50	74.13%	77.61%	75.12%	67.66%	67.66%	73.13%	58.21%

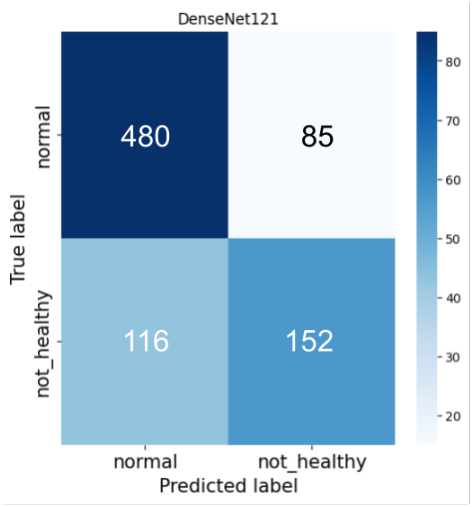
B. Test Set Performance

Using the best configuration selected on the validation split, final evaluation was conducted on a strictly held-out test set. The 3-D DenseNet-121 achieved an overall test accuracy of 75.90%, whereas the 3-D DenseNet-264 achieved 74.70% as shown in Table III. Thus, despite its greater depth and slightly higher peak validation performance in some regions of the search space, DenseNet-264 did not translate this advantage into generalization to the unseen test cohort.

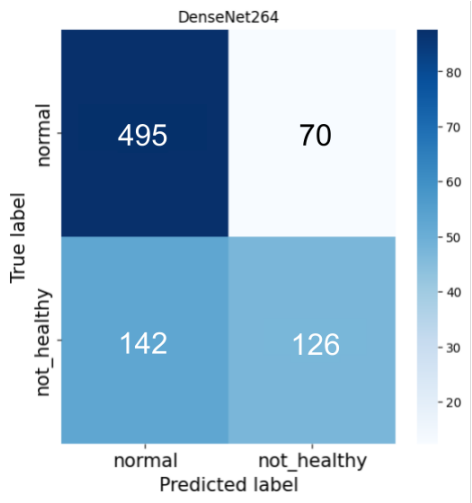
TABLE III. CLASSIFICATION REPORT ON MODEL PERFORMANCES

Models	Accuracy	Precision	Recall	F1 Score
DenseNet121	75.90%	0.75	0.76	0.75
DenseNet264	74.70%	0.74	0.75	0.74

Figure 2 summarizes the class-wise behavior via confusion matrices. Visual inspection shows that DenseNet-264 attains a lower true positive rate (sensitivity) for the Not healthy class but at the cost of a decreased false positive rate, i.e., more Normal scans incorrectly flagged as Not healthy. Similarly, DenseNet-121 attains a higher true positive rate but at the cost of an increased false positive rate.



(a) DenseNet121



(b) DenseNet264

Figure 2. Normalized confusion matrices (in %) for DenseNet121 and DenseNet264.

V. DISCUSSION

The experimental results highlight several important trade-offs between model capacity, optimization, and clinical utility in pediatric brain MRI classification.

First, the accuracy difference between the two architectures on the held-out test set is modest (75.90% for DenseNet-121 vs. 74.70% for DenseNet-264), but it consistently favors the shallower model. Given that both networks were trained on the same curated 0–36 month cohort, this suggests that additional depth and parameters do not automatically translate into better generalization under limited pediatric data. Dense connectivity facilitates feature reuse in both variants, but the larger DenseNet-264 may be more prone to overfitting subtle, cohort-specific patterns unless regularization, data augmentation, or larger training sets further constrain the solution space.

Second, the confusion matrices in Figure 2 show a clinically relevant trade-off. DenseNet-121 attains higher sensitivity for Not healthy scans, which is attractive for a screening or triage role where missing a true abnormality is costly. However, this comes with an increased false positive rate, which could generate more cases for radiologist review that ultimately prove normal. DenseNet-264 attains lower sensitivity for Not healthy

scans.

Third, the hyperparameter sweeps underscore that both architectures are sensitive to learning rate and training duration. DenseNet-121 exhibited robust performance across a band of moderate learning rates (10^{-5} to 10^{-3}) when trained for 30–50 epochs, while DenseNet-264 required a more conservative setting (e.g., 10^{-5}, 40 epochs) to reach its best validation performance. This pattern suggests that future work could benefit from more aggressive regularization (such as stronger augmentation, weight decay, or stochastic depth).

From a workflow perspective, these results support the feasibility of end-to-end 3-D classification of pediatric T2-weighted MRI in the first three years of life. A model with mid-70% test accuracy is not intended to replace expert interpretation but can serve as an assistive tool flagging likely abnormal studies for prioritized review, highlighting challenging cases for second reads, or providing a consistent baseline against which future improvements can be measured.

Future extensions include expanding the dataset to improve generalization, incorporating additional sequences or metadata (e.g., age or protocol information), and validating the models on external cohorts from other scanners and institutions. Integrating interpretability techniques (e.g., 3-D saliency or attribution maps) could further support clinician trust by localizing regions that drive model predictions. Ultimately, the present study establishes a reproducible 3-D DenseNet baseline for pediatric T2 MRI classification and illustrates how architecture depth and hyperparameter choices influence both numerical performance and clinically relevant error profiles.

While direct comparison to radiologist performance was beyond the scope of this study, the model is positioned as an assistive triage mechanism rather than a diagnostic replacement. Future work should incorporate reader studies comparing model output against initial radiologist screening decisions.

A. Limitations

1) Single-Scanner Limitation: All MRIs were acquired from a single scanner and institution. This represents the primary threat to external validity. Scanner-specific acquisition parameters, coil configurations, and reconstruction pipelines may introduce site-dependent biases. Performance may decline when applied to external institutions with different hardware or protocols. Multi-center validation is therefore a critical next step.

2) Heterogeneous "Not Healthy" Label: The "Not healthy" category aggregates multiple pathological conditions, including structural malformations, white-matter abnormalities, and other developmental findings. This coarse binarization simplifies the classification task but obscures subtype-specific performance variation. Future studies should explore multi-class labeling to quantify differential sensitivity across abnormality types.

3) Clinical Utility Threshold: Although mid-70% accuracy demonstrates feasibility, this level of performance is insufficient for autonomous clinical deployment. In real-world triage settings, higher sensitivity—potentially above 90%—would likely be required to ensure minimal missed pathology. Accordingly, the present study should be interpreted as a proof-of-concept baseline rather than a deployable system.

VI. Conclusion and Future Work

This study validates the application of DenseNet architecture variants for classifying pediatric brain MRI, demonstrating the promise of 3-D convolutional networks in enhancing diagnostic accuracy. DenseNet-121 outperformed the deeper DenseNet-264 on the strictly held-out test set, suggesting that increased architectural depth did not improve generalization under limited pediatric data. However, the necessity to address increased false positives remains paramount. These findings contribute an openly documented 3-D pediatric MRI classification baseline that future studies can benchmark against when evaluating larger multi-institutional datasets.

Future investigations will aim to expand our dataset, thereby enhancing model generalization and benchmark comparison. Additionally, exploring hybrid models with attention mechanisms could provide better interpretability and accuracy by focusing computational resources on critical brain regions. Deployment within clinical routines should involve an iterative feedback mechanism from radiologists to facilitate continual learning and system refinement, ultimately supporting efficient and accurate pediatric neurodiagnostic processes.

References

[1] J. M. Zubler et al., "Evidence-informed milestones for developmental surveillance tools", *Pediatrics*, vol. 149, no. 3, Mar. 2022.

[2] S. A. Bélanger and J. Caron, "Evaluation of the child with global developmental delay and intellectual disability", *Paediatr. Child Health*, vol. 23, no. 6, pp. 403–419, Sep. 2018.

[3] A. S. Ali et al., "Magnetic resonance imaging (MRI) evaluation of developmental delay in pediatric patients", *J. Clin. Diagn. Res.*, vol. 9, no. 1, TC21–4, Jan. 2015.

[4] A. M. Pagnozzi et al., "Brain mri before and at term equivalent age predicts motor and cognitive outcomes in very preterm infants", *NeuroImage: Reports*, vol. 5, no. 2, p. 100 262, 2025.

[5] S. M. Jung, "Drug selection for sedation and general anesthesia in children undergoing ambulatory magnetic resonance imaging", *Yeungnam University Journal of Medicine*, vol. 37, no. 3, pp. 159–168, 2020.

[6] F. P. Beaulieu et al., "Sedation and anesthesia for imaging of the infant and neonate—a brief review", *Pediatric Radiology*, vol. 54, no. 10, pp. 1579–1588, 2024.

[7] E. Rula, "Radiology workforce shortage and growing demand something has to give", *Am Coll Radiol*, 2024.

[8] European Society of Radiology (ESR), "What the radiologist should know about artificial intelligence—an esr white paper", *Insights into Imaging*, vol. 10, no. 44, 2019. DOI: 10.1186/s13244-019-0738-2.

[9] M. Annarumma et al., "Automated triaging of adult chest radiographs with deep artificial neural networks", *Radiology*, vol. 291, no. 1, pp. 196–202, 2019. DOI: 10.1148/radiol.2018180921.

[10] X. Liu et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis", *The lancet digital health*, vol. 1, no. 6, e271–e297, 2019.

[11] S. P. Singh et al., "3d deep learning on medical images: A review", *Sensors*, vol. 20, no. 18, p. 5097, 2020.

[12] A. Bhatia, F. Khalvati, and B. B. Ertl-Wagner, "Artificial intelligence in the future landscape of pediatric neuroradiology: Opportunities and challenges", *AJNR American Journal of Neuroradiology*, vol. 45, no. 5, pp. 549–553, 2024. DOI: 10.3174/ajnr.A8086.

[13] G. Litjens et al., "A survey on deep learning in medical image analysis", *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[14] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", in *International conference on medical image computing and computer-assisted intervention*, Springer, 2016, pp. 424–432.

[15] A. E. Ilesanmi, T. O. Ilesanmi, and B. O. Ajayi, "Reviewing 3d convolutional neural network approaches for medical image segmentation", *Heliyon*, vol. 10, no. 6, 2024.

[16] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)", *IEEE Trans. Med. Imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[17] S. P. Singh et al., "3d deep learning on medical images: A review", *Sensors*, vol. 20, no. 18, p. 5097, 2020. DOI: 10.3390/s20185097.

[18] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation", in *MICCAI*, ser. Lecture Notes in Computer Science, vol. 9901, Springer, 2016, pp. 424–432. DOI: 10.1007/978-3-319-46723-8_49.

[19] K. Kamnitsas et al., "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation", *Medical Image Analysis*, vol. 36, pp. 61–78, 2017. DOI: 10.1016/j.media.2016.10.004.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks", in *CVPR*, 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.